DOCUMENT RESUME

ED 307 306 TM 013 360

AUTHOR Jaeger, Richard M.
TITLE Selection of Judges for Standard Setting: What Kinds? How Many?
PUB DATE Mar 89
NOTE 17p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, March 28-30, 1989).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS College Entrance Examinations; Cutting Scores; *Difficulty Level; Estimation (Mathematics); Evaluation Criteria; *Evaluators; Graduation Requirements; Higher Education; High School Students; Interrater Reliability; Latent Trait Theory; Licensing Examinations (Professions); Minimum Competency Testing; Qualifications; Secondary Education; *Selection; *Standardized Tests; Teacher Certification; Teacher Education; Test Items
IDENTIFIERS *Experts; *Standard Setting

ABSTRACT

Criteria for the selection of judges (evaluators) for setting item-based standards involved in tests for which cutting scores must be established are investigated. Focus is on cases in which test standards are based on specialists' judgments concerning the difficulty of test items in tests used to determine who will be awarded a diploma, admitted to a program, certified as competent, or licensed to practice. The literature on the nature of expertise describe the qualifications that are required of judges assigned to apply item-based standard-setting procedures. Qualifications for judges of items for tests used to select applicants to teacher education programs, grant initial teacher certificates, and grant high school diplomas are briefly outlined. Post hoc analyses of judges are preferable to competency testing of judges. An item response model outlined by Van der Linden (1982) and an alternative approach involving interrater comparisons outlined by Jaeger (1988) are offered as possible means of post hoc analysis. Problems associated with determination of a sufficient number of judges, including issues involving standard errors of mean and measurement, are outlined. In general, judges should be selected through procedures that permit generalization of their collective recommendations to well-defined populations. The numbers of judges selected should be sufficient to provide precise estimation of the standard that would be recommended by an entire population of judges. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED307306

Selection of Judges for Standard Setting: What Kinds? How Many?

Richard M. Jaeger
Center for Educational Research and Evaluation
University of North Carolina at Greensboro

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

RICHARD M. JAEGER

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Presented in a symposium on Practical Issues in Conducting a Standard-Setting Study at the 1989 Annual Meeting of the National Council on Measurement in Education, San Francisco, March 28-30.

TM013360

BEST COPY AVAILABLE

Overview

When testing is used to determine who will be awarded a diploma, admitted to a program, certified as competent, or licensed to practice, it is necessary to determine the minimum level of test performance associated with a positive selection decision. In some situations, the minimum required test performance is specified by tradition or law (e.g., 70 percent correct is the "passing score") or is based on the normative performance of examinees (e.g., the 50th percentile of the test score distribution). Although such minimums (termed test standards or cutoff scores) are based on subjective judgment, it is usually the judgment of a few policy makers whose authority to establish test standards derives from position rather than qualification. This paper excludes such situations, and considers cases in which test standards are based on specialists' judgments concerning the difficulty of test items. Such specialists are often referred to as "experts." In contrast to policy-makers, experts' authority to recommend test standards derives solely from their qualifications.

Three practical questions arise in the situations considered here: Who should make judgments about test standards; i.e., who is an expert? How should these judges be selected? How many judges should be selected? This paper examines these questions, describes some analytically-derived bases for their answers, and provides some advice on the judgments they require when scientific analysis fails to inform.

What Kinds of Judges?

Unfortunately, the *Test Standards* (American Psychological Association, 1985) pay little attention to the selection of judges for setting standards on tests. Only Standard 6.9 addresses the issue of judges' qualifications directly, and then solely in terms of documentation:

"When cut scores are based primarily on professional judgment, the qualifications of the judges should be documented" (p. 43).

We must therefore appeal to other sources when considering the question: Who should make judgments about test standards?

The most frequently used item-based standard-setting methods require complex judgments and the ability to make decisions rapidly. Reasonable results can be expected only if the judges called upon to use these methods are highly knowledgeable of the domain in which decisions are required.

The judgments required by Angoff's (1971) and Nedelsky's (1954) procedures (standard-setting methods that are claimed by Smith and Smith (1988) to be used more frequently than any others) might be described as a mixture of judgmental predictive validation and judgmental item analysis. Angoff's method, for example, requires the conceptualization of a minimally competent examinee and, for such an examinee, prediction of the probability that each item on the test would be answered correctly. If an examination is to be used for purposes of admission screening or some other selection purpose, a minimally competent examinee is defined in terms of success in the activity for which selection is sought. When a test is used to screen applicants to a teacher education program, for example, a minimally competent examinee is a person who would barely succeed in such a program. When asked to recommend a standard for a test used to screen applicants to a teacher education program, a judge must therefore be capable of rendering accurate judgments on the demands of typical teacher education programs, on the abilities of students who are barely able to succeed in such programs, and on the difficulties of test items that purportedly assess such abilities. Such judgments require substantial knowledge concerning these topics; knowledge that is rapidly accessible and readily integrated. In short, judges must be experts in the domain assessed by

the test and in the roles sought by successful examinees. We might therefore be guided in formulating qualifications for such judges by examining the concept of expertise and the definition of "expert."

Documenting the results of their own research as well as that of others, Chi, Glaser and Farr (1988) describe experts as:

(1) excelling mainly in their own domains. For example, Minsky and Papert (1974) note that "A very intelligent person might be that way because of specific local features of his knowledge-organizing knowledge rather than because of global qualities of his 'thinking' " (p. 59);

(2) able to perceive large meaningful patterns in their domain of expertise. This ability does not reflect generally superior perceptual ability, but rather possession of an organized knowledge base;

(3) able to perform rapidly in their domain of expertise. Experts are faster than novices at performing skills in their domain and they quickly solve problems with little error. Experts' speed is based on practice and resulting automaticity, and also the capability of arriving at a solution to a problem without extensive search of the knowledge base;

(4) experts see and represent a problem in their domain at a deeper (more principled) level than do novices. Novices tend to represent problems at a superficial level. Both experts and novices divide a domain of knowledge into conceptual categories, but the experts' categories are semantically or principle-based, whereas the categories of novices are syntactically or surface-feature oriented;

(5) experts spend time analyzing a problem qualitatively. Experts build a mental representation of a problem from which they infer relationships

that define the problem situation and then they add constraints. Adding constraints reduces the "search space" for solutions;

(6) experts have strong self-monitoring skills. Experts seem to be more aware than novices of when they make errors, why they fail to comprehend, and when they need to check their solutions;

(7) experts are more accurate than novices at judging problem difficulty (Chi, Glaser & Rees, 1982). The superior monitoring skills and self-knowledge of experts reflect their greater domain knowledge as well as a different representation of that knowledge. Novices use surface features to sort problems on difficulty. They also use facts outside the problem (e.g., "I've never seen a problem like that before."). Experts use underlying principles from the knowledge domain that are required to solve a problem as a basis for sorting problems on difficulty (Glaser, 1987);

(8) expertise lies more in an elaborated semantic memory than in a general reasoning process (Posner, 1988).

In sum, the literature on the nature of expertise describes quite well, the qualifications that we require of judges when we have them apply an item-based standard-setting procedure. We must ask, then, where do we find such experts, and how can we know we have found them?

In his pursuit of the "expert pedagogue," Berliner (1986) concluded that expertise and experience were so completely confounded that it was impossible to differentiate them. He used the two interchangeably. This suggests that our expert judges should be well experienced in the domains of expertise we demand of them. In the case of setting standards on tests used to screen applicants to teacher education programs, for example, expert judges should have experience with the curricula of teacher education programs, with

the knowledge demands imposed by such programs, and with the capabilities of students who succeed in such programs.

We can speculate on the backgrounds of judges most likely to satisfy the definition of "expert" in several of the most common standard-setting applications:

Standards for admission to teacher education programs. Those most likely to be expert judges of appropriate test standards for admission to teacher education programs are experienced faculty members in colleges or universities who provide instruction in teacher education programs, and classroom teachers who are relatively recent graduates of the teacher education programs for which admission is sought.

Standards for granting initial certification to teach. Judges used for this standard-setting task should know the prerequisite knowledge requirements of the job of beginning teacher, and the capabilities of those who occupy the role of beginning teacher. Those most likely to possess such knowledge are teacher educators who supervise student teachers, and fully certified practicing teachers who either supervise beginning teachers or who have a few years of successful experience but are not so far beyond their initial year of teaching that they have forgotten the demands their job imposed during that initial year.

Standards for granting high school diplomas. Judges used for this standard-setting purpose should know the knowledge requirements of entry-level, post-high school jobs or freshman courses in colleges and universities, assuming the purpose of high school graduation tests is to ensure that high school graduates possess knowledge sufficient to enter the labor force or enter a postsecondary education program. Judges most likely to possess this kind of expertise include directors of apprenticeship programs for craft unions, personnel directors of service-oriented companies that hire large numbers of

recent high school graduates, college and university admissions officers and college and university faculty members who teach freshman courses.

These nominations are admittedly speculative. Their unifying characteristic is the extensive direct experience that is likely to be brought to the judgment task by members of the nominated groups. Clearly, not all directors of craft union apprenticeship programs will have the rich, differentiated store of knowledge that is essential to expertise in the domain of qualification for high school graduation. Nor will all experienced teachers be classifiable as experts on the knowledge demands imposed on successful beginning teachers. Methods for identifying members of these groups who have legitimate claim to expertise are sorely needed.

It is tempting to suggest the use of competency tests for those who would recommend standards of competence. But such tests would be impractical, and even if practical, would impose standard-setting problems of their own. Identification of judges with sufficient expertise might be far more readily accomplished through *post hoc* analyses of judges' recommendations than through direct examination of their knowledge structures. Several promising analytic procedures have been proposed.

Van der Linden (1982) defined a component of error variance in standard setting what he termed "intrajudge inconsistency" (p. 296). He developed an index that reflects the degree of consistency exhibited by individual judges when they provide the kinds of judgments imposed by the Angoff (1971) and Nedelsky (1954) standard-setting procedures. The validity of the Van der Linden index depends on the goodness of fit of the items that compose a test to an item response theory (IRT) model. The index is based on the supposition that judges have in mind, a minimally competent examinee with a specified level of ability on the construct assessed by the test. The

consistency index of judges' recommendations is computed by examining deviations from estimates that would result if the IRT model were applied to the item performances of an examinee of the specified ability. To the extent that judges with greater expertise are more consistent in their recommendations, the Van der Linden index might be useful in identifying judges who were experts and judges who were relative novices.

An alternative approach to identifying judges whose patterns of judgments on test items are aberrant was proposed by Jaeger (1988). The methodology Jaeger used was first suggested by Sato (1975) to identify examinees whose performances on sequences of test items suggested that their knowledge base was different from those of most examinees and, by implication, that the construct measured by the examination was different for them than for most examinees. When applied to selection of judges used for standard setting, this "caution index" approach assumes that most selected judges are experts and, by virtue of their common apperception and organization of the knowledge domain essential to their judgment task, will be consistent in their patterns of item recommendations. Judges whose patterns of recommendations are inconsistent with those of the majority are thus identified as relative novices.

The notion of consensus among experts, and the use of a consistency index to identify those whose judgments should be afforded greater credence has also been proposed in the field of quantitative anthropology by Batchelder and Romney (1986), Weller (1984), and Romney, Weller and Batchelder (1986). The utility of their methods for identifying expert judges in standard-setting applications remains to be demonstrated.

How Many Judges?

The question "How many judges should be used in a standard-setting study?" can be viewed as a special case of "How large a sample do I need?" This is the second question that neophyte researchers invariably ask of statisticians or major professors . The pat answer is "As many as you can get." But this answer is sometimes unreasonable because "as many as you can get" might well be insufficient or wasteful for any practical purpose.

If judgments could be obtained from an entire population of judges, we would obviously consider the resulting test standard to be more trustworthy than one based on the judgments of a small sample. Many standard-setting procedures define a test standard as the mean of the standards recommended by all sampled judges. For these procedures, we can consider the mean standard that would be recommended by an entire population of qualified judges to be a population parameter. The mean of the standards recommended by a sample of judges can, likewise, be regarded as an estimate of this population parameter. The difference between a sample estimate and the parameter to be estimated is termed an estimation error. Estimation errors will vary across samples of a given size. The number of judges sampled should be sufficiently large that the magnitude of most estimation errors will be tolerable.

The magnitude of estimation errors can be characterized by the standard error of the mean. In the current application, the standard error of the mean is the standard deviation of the distribution of mean recommended test standards for samples of judges of a given size. If judges are sampled randomly, independently, and with equal probability, the standard error of the mean can be estimated by the standard deviation of the test standards recommended by individual judges, divided by the square root of the number of judges sampled:

Standard Error of the Mean = (Standard Deviation)/$\sqrt{(n)}$ (Eq. 1)

where n is the number of judges in the sample.

If the number of judges sampled is sufficiently large, we know by the Central Limit Theorem (Glass and Hopkins, 1984; 184-190) that the estimation error will not exceed two standard errors of the mean for about 95 percent of the samples of that size that could be selected. This knowledge can be used to determine the number of judges that must be sampled to provide reasonable assurance that the estimation error will not exceed a tolerable value.

To estimate the standard error of the mean using Equation (1) we must know, or have an estimate of, the standard deviation of test standards recommended by individual judges. Although constant in a particular study, this standard deviation will vary across studies, depending on such factors as the type of judge used (Jaeger, 1982), the standard-setting procedure used (Cross, Impara, Frary & Jaeger, 1984), and the length of the test for which standards are sought (Jaeger, in press). There is some limited evidence that the standard deviation of judges' recommended test standards will be larger for longer tests than for shorter tests. It is likely that the training judges receive also will influence the standard deviation of their recommended test standards; one would expect effective training to reduce this standard deviation.

Data on the standard deviation of recommended test standards when judges used Angoff's (1971) method were gleaned from five empirical studies that used a total of nine tests (Brennan & Lockwood, 1980; Cross, Impara, Frary & Jaeger, 1984; Mills & Melican, 1987; Melican & Thomas, 1984; Smith & Smith, 1988). The tests used in these studies ranged in length from 34 items to 150 items and all were in four-option, multiple choice format. The smallest observed standard deviation of recommended test standards was 0.79 items (for a 34-item test) and the largest was 29.03 items (for a 150-item test). The

median of the standard deviations was 4.65 items (for a 110-item test). We will use the median of the standard deviations of recommended test standards to illustrate determination of the number of judges who would have to be sampled in order to reduce the standard error of the mean to an acceptable level, or equivalently, to reduce a 95 percent confidence interval on the population mean recommended test standard to an acceptable width.

Assume for the moment, that 15 judges had been sampled, a value not unlike those commonly encountered in published standard-setting studies (e.g., Cross, et al., 1984; Smith & Smith, 1988). In that case, the estimated standard error of the mean recommended test standard would be:

Standard Error of the Mean = $(4.65)/\sqrt{(15)} = 1.20$ test items.

Correspondingly, an approximate 95 percent confidence interval on the population mean recommended test standard would be approximately 4.80 test items wide or, put another way, approximately 95 percent of simple random samples of 15 judges would yield mean recommended test standards with estimation errors that did not exceed 2.40 test items.

Whether a standard error of the mean equal to 1.20 test items is acceptable is a matter of judgment. One could argue that no examinee (or very few examinees) with a test score that was within one point of the test standard that would have been recommended by the total population of judges should be misclassified (as having passed or failed the test), just due to estimation errors that arise from sampling of judges. Under that argument, the standard error of the mean recommended test standard should not exceed 0.5 test items, a value necessary to reduce the estimation error to no more than one test item for 95 percent of the samples of judges that might be drawn (assuming a fixed sample size). We can rearrange Equation (1) to solve for the required sample size, as follows:

n = (Standard Deviation/Desired Standard Error of the Mean)^2 Eq. (2)

For the data at hand,

$n = (4.65/0.5)^2 = 86.49$,

which indicates that 87 judges would have to be sampled to ensure an estimation error not exceeding one test item with 95 percent probability.

A standard-setting study with a sample of 87 judges might be unacceptably expensive or impractical, although larger studies have been conducted (cf. Jaeger, 1982; Jaeger & Busch, 1984). An alternative criterion for determining an acceptable size for the standard error of the mean recommended test standard can be developed by considering other sources of error that arise when an examinee is classified (as having passed or failed) on the basis of an observed test score. A major source of error derives from the less-than-perfect reliability of certification and competency tests. This error source can be characterized by the standard error of measurement of the test.

If errors due to the unreliability of the test and errors due to sampling of judges are considered to be independent (a plausible assumption), a standard error that reflects both sources is given by:

Standard error = $\sqrt{(\text{Standard Error of the Mean}^2 + \text{Standard Error of Measurement}^2)}$ (Eq. 3)

Equation (3) can be used to show that the overall standard error is only three percent larger than the standard error of measurement, provided the standard error of the mean is no larger than one-fourth the size of the standard error of measurement of the test. If the standard error of the mean were half the size of the standard error of measurement, the overall standard error would only be 12 percent larger than one resulting from measurement error alone.

One might thus argue that a sufficiently large sample of judges is one that will result in a standard error of the mean recommended test score that was no

larger than one-fourth the standard error of measurement of the test for which a standard is sought. Fewer judges might be required under this criterion. For example, the *NTE Programs Test Analysis Booklet -- Core Battery* (1984) reports the standard error of measurement of the National Teacher Examinations General Knowledge Test as 5.2 items. If the standard error of the mean recommended test standard was one-fourth of this value, 1.3 items, the overall standard error would equal 5.36 items, an increase of only 3.08 percent. The number of judges required to produce a standard error of the mean recommended test standard equal to 1.3 test items would be

$$n = (4.65/1.3)^2 = 12.79,$$

which rounds up to 13 judges, assuming the standard deviation of judges' recommended test standards equals 4.65, the median for nine studies that used Angoff's standard-setting procedure, as reported earlier.

Some researchers have suggested that test standards be based on the median of the distribution of judges' recommendations, rather than the mean of that distribution. This suggestion is designed to reduce the effect of the most extreme standard-setting recommendations (such as requiring examinees to answer every item correctly in order to pass a test). It is well known that the sample median is a less stable statistic than the sample mean. When samples are selected from a population with a normal distribution of scores, the standard error of the sample median will be 1.253 times as large as the standard error of the mean. If the standard error of the median or the width of a 95 percent confidence interval around the sample median are to be controlled, a slightly larger sample of judges will be required than would be needed to produce an equivalent standard error of the mean. In the previous example, the standard error of the sample median would be (1.3) * (1.253) = 1.629. When combined with a standard error of measurement equal to 5.2 in Equation (3), the overall

standard error becomes 5.45, an increase of 4.79 percent over the standard error of measurement alone. The overall standard error could be held to the previously considered 5.36 by reducing the standard error of the median recommended test standard to 1.3. By Equation (2), this would require sampling 20.09 or, rounding up, 21 judges.

Summary

The tasks imposed by item-based standard-setting procedures require that judges have ready access to richly organized knowledge bases. The models of expertise distilled by Chi, Glaser and Farr (1988) and their colleagues describe well, the expectations effective judges must satisfy. Berliner's (1986) equating of expertise and experience suggests limited operational guidelines for identifying classes of judges who are likely to meet the demands imposed by item-based standard-setting procedures. It might be possible to identify individuals within those classes who meet the requirements of expertise by using statistical procedures that examine the internal consistency of a judge's recommendations or the congruence of a judge's recommendations with those of a majority of presumed experts.

Judges should be selected through procedures that permit generalization of their collective recommendations to well-defined populations. The literature on sampling from finite populations includes such procedures. The numbers of judges selected should be sufficient to provide precise estimation of the standard that would be recommended by an entire population of judges. The size of the sample of judges should be such that the standard error of the mean or median recommended standard is small, compared to the standard error of measurement of the test for which a standard is sought.

References

American Psychological Association, American Educational Research Association, National Council on Measurement in Education (1985). *Standards for educational and psychological testing.* Washington, DC: Author.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.). *Educational measurement* (2nd. ed.). Washington, DC: American Council on Education.

Batchelder, W. H. & Romney, A. K. (1986). The statistical analysis of a general Condorcet model for dichotomous choice situations. In B. Grofman & G. Owen (Eds.). *Information pooling and group decision making.* Hartford, Connecticut: JAI Press.

Brennan, R. D. & Lockwood, R. E. (1980). A comparison of Nedelsky and Angoff cutting score procedures using generalizability theory. *Applied Psychological Measurement, 4,* 219-240.

Chi, M., Glaser, R. & Farr, M. (Eds.). (1988). *The nature of expertise.* Hillsdale, NJ: Lawrence Erlbaum.

Chi, M., Glaser, R., & Rees, F. (1982). Expertise in problem solving. In R. Sternberg (Ed.). *Advances in the psychology of human intelligence.* Vol. I, pp. 17-76. Hillsdale, NJ: Lawrence Erlbaum.

Cross, L. H., Impara, J. C., Frary, R. B. & Jaeger, R. M. (1984). A comparison of three methods for establishing minimum standards on the National Teacher Examinations. *Journal of Educational Measurement, 21,* 113-130.

Educational Testing Service (1984,February). *NTE programs test analysis booklet -- core battery.* Unpublished statistical report SR-84-19. Princeton, NJ: Author

Glaser, R. (1987). Thoughts on expertise. In C. Schooler & W. Schaie (Eds.). *Cognitive functioning and social structure over the life course.* (pp. 81-94). Norwood, NJ: Ablex.

Glass, G. V & Hopkins, K. D. (1984). *Statistical methods in education and psychology.* (2nd. ed.). Englewood Cliffs, NJ: Prentice-Hall.

Jaeger, R. M. (1982). An iterative structured judgment process for establishing standards on competency tests: Theory and application. *Educational Evaluation and Policy Analysis, 4,* 461-475.

Jaeger, R. M. (1988). Use and effect of caution indices in detecting aberrant patterns of standard-setting judgments. *Applied Measurement in Education, 1,* 17-31.

Jaeger, R. M. (in press). Setting standards on teacher certification tests. In J. Millman & L. Darling-Hammond (Eds.). *Handbook for the evaluation of elementary and secondary school teachers.* Newbury Park, CA: Sage Publications.

Melican, G. C. & Thomas, N. (1984, April). Identification of items that are hard to rate accurately using Angoff's standard setting method. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Mills, C. N. &Melican, G. C. (1987). *A preliminary investigation of three compromise methods for establishing cut-off scores.* Research Report RR-87-14. Princeton, NJ: Educational Testing Service.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14,* 3-19.

Posner, M. I. (1988). Introduction: What is it to be an expert? In Chi, M., Glaser, R. & Farr, M. (Eds.). *The nature of expertise.* Hillsdale, NJ: Lawrence Erlbaum.

Romney, A. K., Weller, S. C. & Batchelder, W. H. (1986). Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist, 88,* 313-336.

Sato, T. (1975). [The construction and interpretation of S-P tables.] Tokyo: Meiji Tosho.

Van der Linden, W. (1982). A latent trait method for determining intrajudge inconsistency in the Angoff and Nedelsky techniques of standard setting. *Journal of Educational Measurement, 4,* 295-305.

Weller, S. C. (1984). Consistency and consensus among informants: Disease concept in a rural Mexican town. *American Anthropologist, 86,* 966-975.